Attività formativa: elaborato

Un'applicazione per la raccolta di giudizi di qualità di pagine web

Laureando: Roberto Compostella

Relatore: Prof. Luca Pretto
Correlatore: Prof. Enoch Peserico

Corso di Laurea in Ingegneria Informatica

30/09/2010

Anno Accademico 2009/2010

Sommario

I primi sistemi di Information Retrieval operavano su collezioni di documenti di qualità omogenea scritti da autori affidabili e autorevoli. Il loro unico scopo era quello di restituire i documenti più rilevanti in base alle interrogazioni formulate dagli utenti. Nel web moderno invece, a causa dell'eterogeneità della qualità delle pagine web, l'applicazione delle tecniche di Information Retrieval si è rilevata poco efficace; di qui la necessità di ideare algoritmi in grado di selezionare le pagine web anche in base alla loro qualità. I principali sono gli algoritmi di link analysis che cercano di inferire la qualità delle pagine web dalla struttura topologica del grafo associato al web.

Il lavoro descritto in questa relazione è stato svolto all'interno di un progetto che ha lo scopo di valutare l'effettiva efficacia degli algoritmi di link analysis nell'individuare le pagine web di maggiore qualità. Il nostro lavoro è consistito nel realizzare un applicativo web e di studiare una tecnica per campionare le pagine web restituite da dieci interrogazioni scelte ad hoc.

L'applicativo web permette, dopo aver effettuato la registrazione o il login, di raccogliere i giudizi di qualità di pagine web secondo alcuni parametri.

Per quanto riguarda il campionamento è stato adottato il campionamento a due stadi, nel quale prima si divide la popolazione in G grappoli in base alla rilevanza del documento e poi si selezionano da ciascun grappolo le singolo unità statistiche attraverso il campionamento sistematico.

Indice

1. Introduzione

Presso l'Università degli Studi di Padova si sta svolgendo un progetto di ricerca volto a valutare la bontà degli algoritmi di link analysis. Tali algoritmi servono per selezionare le pagine web in base alla loro qualità partendo da un grafo orientato che rappresenta il web italiano. Lo scopo del progetto è mettere a confronto gli algoritmi di link analysis, come PageRank [3] e HITS [4], con i giudizi di qualità espressi da utenti umani su una popolazione di URL mediante un'applicazione web per dedurre se i punteggi attribuiti da tali algoritmi sono in relazione con la qualità effettiva dei documenti.

Hanno seguito questo progetto alcuni studenti del terzo anno di Ingegneria Informatica, divisi in due gruppi con finalità diverse. Uno si è occupato di effettuare il crawling del web italiano con il relativo grafo per poi eseguire gli algoritmi di link analysis sui vari documenti. L'altro ha sviluppato un applicativo per raccogliere i giudizi di rilevanza e qualità delle pagine web e ha sottoposto ai motori di ricerca 10 interrogazioni, campionando e scaricando le pagine restituite in base alla rilevanza.

Più precisamente abbiamo realizzato un'applicazione web, il cui prototipo era stato sviluppato da altri studenti dell'Università di Padova, e studiato un metodo di campionamento per selezionare alcune pagine tra tutte quelle restituite dai motori di ricerca.

Il motivo per cui ci siamo occupati di questo progetto è dovuto al fatto che nel web attuale ci sono moltissimi documenti di scarsa qualità creati da persone non autorevoli e\o non affidabili. Dobbiamo quindi capire se gli algoritmi oggi utilizzati possono adempiere al compito per cui vengono utilizzati. Ci siamo limitati al solo web italiano in quanto:

- è impossibile valutare tutto il web essendo molto vasto;
- il web italiano è abbastanza isolato dato che si sono pochi link uscenti ed è ragionevole pensare che la struttura sia ricorsivamente simile a tutto il web;
- le pagine scritte in italiano sono più facilmente valutabili dagli utenti.

2. Elementi di Information Retrieval

L'Information Retrieval (letteralmente recupero dell'informazione) è la branca dell'informatica che si occupa di

> *recuperare materiale (di solito documenti) non strutturato (di solito testo) che soddisfa un bisogno informativo all'interno di grandi collezioni (di solito memorizzate su computer)* [1].

Nella ricerca sul web, il sistema di Information Retrieval deve consentire la ricerca fra miliardi di documenti archiviati su milioni di computer, con efficienza ed efficacia. Per far ciò il sistema deve pre-elaborare i documenti della collezione. In estrema sintesi, quando un documento viene inserito nella collezione, il sistema lo analizza e ne estrae i termini, producendo una matrice di incidenza (figura 2.1), dove $M(t, d) = 1$ se e solo se il documento d contiene il termine t. L'informazione contenuta nella matrice di incidenza viene poi organizzata come posting list (figura 2.2), cioè una lista ordinata di termini che fungono da chiave e ogni termine punta alla lista dei documenti che lo contengono.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Figura 2.1.: Esempio di matrice di incidenza

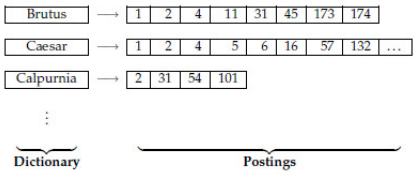

Figura 2.2.: Esempio di posting list relativa alla matrice in figura 2.1

Una volta ottenuto l'indice serve un algoritmo che dia il punteggio di rilevanza di ogni documento in risposta ad un'interrogazione specifica. Ciò viene tipicamente ottenuto tenendo

conto della frequenza dei termini presenti nell'interrogazione nei documenti e misurando la similarità tra interrogazione e documenti [1].

Tutte queste tecniche hanno lo scopo di recuperare, nell'insieme dei documenti restituiti, molti rilevanti e pochi non rilevanti. Esistono due misure che rappresentato l'efficacia nel reperimento delle informazioni:

Precisione È il rapporto tra il numero di documenti rilevanti recuperati e il numero di tutti i documenti recuperati.

Richiamo È il rapporto tra il numero di documenti rilevanti recuperati e il numero di tutti i documenti rilevanti disponibili nella collezione considerata.

$$P = \frac{\text{reperiti} \cap \text{rilevanti}}{\text{reperiti}} \qquad R = \frac{\text{reperiti} \cap \text{rilevanti}}{\text{rilevanti}}$$

Se ipoteticamente venissero recuperati tutti e soli i documenti rilevanti entrambe le misure sarebbero pari ad 1. Nella realtà i motori di ricerca devono trovare un buon compromesso fra le due misure. Nella pratica, se si vuole una precisione alta non vengono reperiti tutti i documenti rilevanti abbassando quindi il richiamo; viceversa, se si vuole un richiamo alto vengono reperiti anche documenti non rilevanti abbassando la precisione.

Dato le grandi dimensioni di collezioni su cui i sistemi di Information Retrieval operano sono necessarie delle tecniche per valutare le prestazioni di tali sistemi. Una di queste tecniche è il pooling. Attraverso questa tecnica si selezionano solo i primi k documenti restituiti da diversi sistemi di Information Retrieval, normalmente quelli da valutare [1]. Il pooling è efficace dal momento che molti dei documenti pertinenti ad un argomento cercato verranno visualizzati dai sistemi di recupero nella prima parte dei risultati e di conseguenza tali documenti saranno giudicati e utilizzati per valutare in maniera efficace le prestazioni dei sistemi di raccolta e indicizzazione.

A causa dell'eterogeneità del web sono stati creati degli algoritmi per determinare la qualità dei documenti; i più significativi sono gli algoritmi di link analysis, che tentano di attribuire un punteggio di qualità partendo dalla struttura topologica del grafo del web. I principali algoritmi di link analysis sono PageRank [3], HITS [4] e SALSA [5].

Il nostro scopo è quello di valutare gli algoritmi di link analysis confrontando i punteggi ottenuti da tali algoritmi con i giudizi di qualità dati dagli utenti alle pagine web.

3. Realizzazione di un applicativo per la raccolta di giudizi di qualità di pagine web

Per raccogliere i giudizi di qualità di pagine web abbiamo progettato e realizzato un applicativo in php che, dopo essersi registrati o aver effettuato il login, permette di esprimere dei voti secondo alcuni parametri.

Alla base dell'applicativo c'è una base di dati, documentata nell'appendice B.

3.1. Home page

L'home page è molto semplice, come si può vedere dalla figura 3.1: nella parte sinistra si può effettuare il login, registrarsi o recuperare la password tramite l'indirizzo e-mail fino ad un massimo di 3 volte al giorno; mentre a destra c'è una breve descrizione dell'applicativo e della sua funzione.

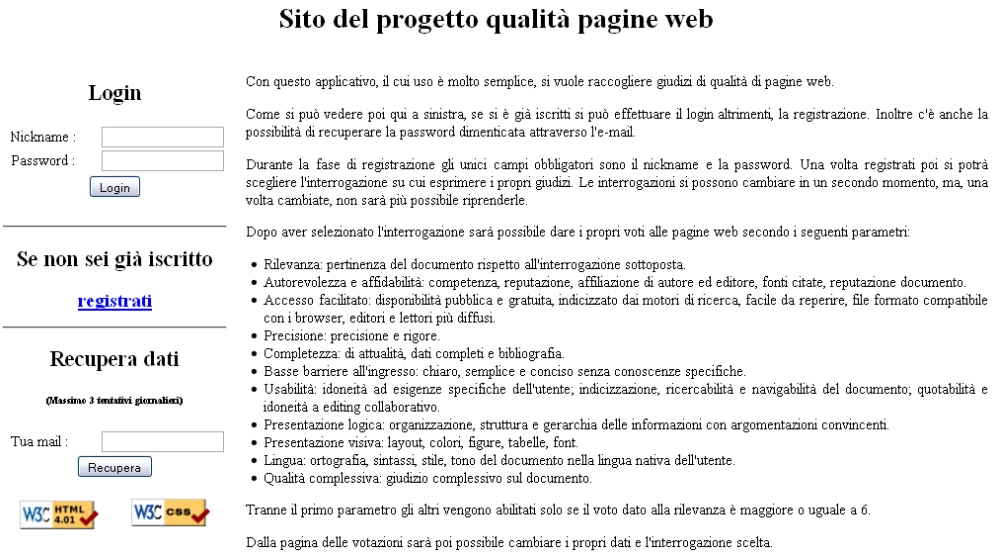

Figura 3.1.: Screenshot dell'home page

3.2. Iscrizione

La pagina per effettuare l'iscrizione è formata da un singolo form realizzato grazie agli SpryAssets (http://labs.adobe.com/technologies/spry/home.html), script che permettono di verificare se i valori inseriti sono nel formato corretto, ad esempio e-mail e data, e per i campi obbligatori se è stato effettivamente inserito un valore.

Oltre ai campi obbligatori nickname e password ci sono i campi sesso, nazionalità, data di nascita, professione, e-mail e matricola che, in base ai dati inseriti, potranno successivamente servire per effettuare un'analisi più accurata.

Simile a quella appena descritta c'è poi una pagina per modificare successivamente i dati inseriti.

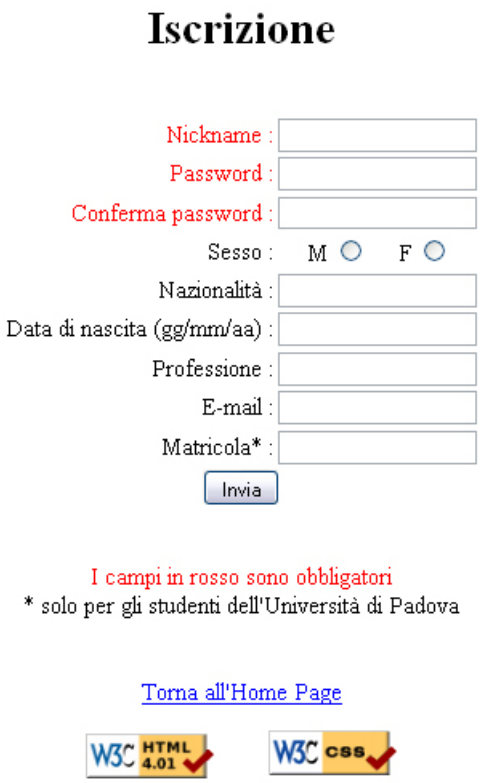

Figura 3.2.: Screenshot della pagina per l'iscrizione

3.3. Selezione dell'interrogazione

In questa pagina, come mostrato in figura 3.3, è possibile selezionare e successivamente cambiare l'interrogazione su cui effettuare le votazioni.

Le interrogazioni sono elencate in ordine crescente rispetto al numero di utenti che le hanno selezionate per invogliare un utente a selezionare le prime, così da mantenere sempre lo stesso numero di utenti che votano in ogni interrogazione. Inoltre non è possibile riprendere un'interrogazione interrotta.

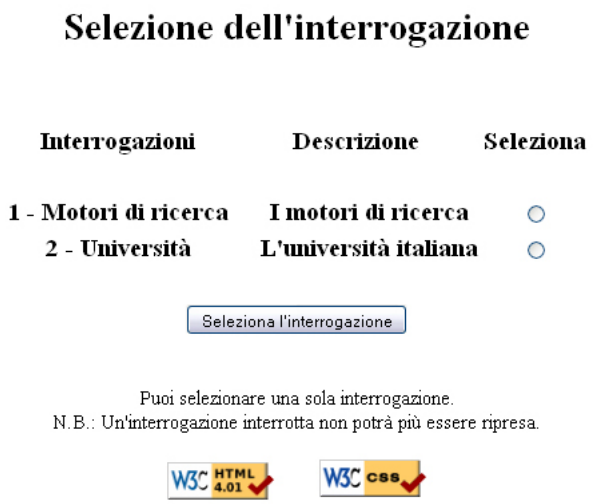

Figura 3.3.: Screenshot della pagina di selezione dell'interrogazione

3.4. Votazioni

Questa è la pagina più importante dell'applicativo. Come si vede in figura 3.4, qui si effettuano le votazioni alle varie pagine web secondo i seguenti parametri:

Rilevanza Pertinenza del documento rispetto all'interrogazione sottoposta.

Autorevolezza e affidabilità Competenza, reputazione, affiliazione di autore ed editore, fonti citate, reputazione documento.

Accesso facilitato Disponibilità pubblica e gratuita, indicizzato dai motori di ricerca, facile da reperire, file formato compatibile con i browser, editori e lettori più diffusi.

Precisione Precisione e rigore.

Completezza Di attualità, dati completi e bibliografia.

Basse barriere all'ingresso Chiaro, semplice e conciso senza conoscenze specifiche.

Usabilità Idoneità ad esigenze specifiche dell'utente; indicizzazione, ricercabilità e navigabilità interna al documento; quotabilità e idoneità a editing collaborativo.

Presentazione logica Organizzazione, struttura e gerarchia delle informazioni con argomentazioni convincenti.

Presentazione visiva Layout, colori, figure, tabelle, font.

Lingua Ortografia, sintassi, stile, tono del documento nella lingua nativa dell'utente.

Qualità complessiva Giudizio complessivo sul documento.

Tranne il primo parametro gli altri vengono abilitati solo se il voto dato alla rilevanza è maggiore o uguale a 6.

Ogni volta che viene inserito un voto l'applicativo provvede a inserirlo immediatamente nella base di dati anche senza premere sui bottoni indietro o avanti.

Per evitare che gli utenti diano i voti senza pensarci le pagine web vengono votate da più utenti. In questo modo si può desumere con più sicurezza se una pagina è rilevante o meno confrontando i vari voti espressi tenendo conto di media e varianza.

In questa pagine ci sono poi dei link per cambiare i propri dati, cambiare l'interrogazione scelta ed effettuare il logout.

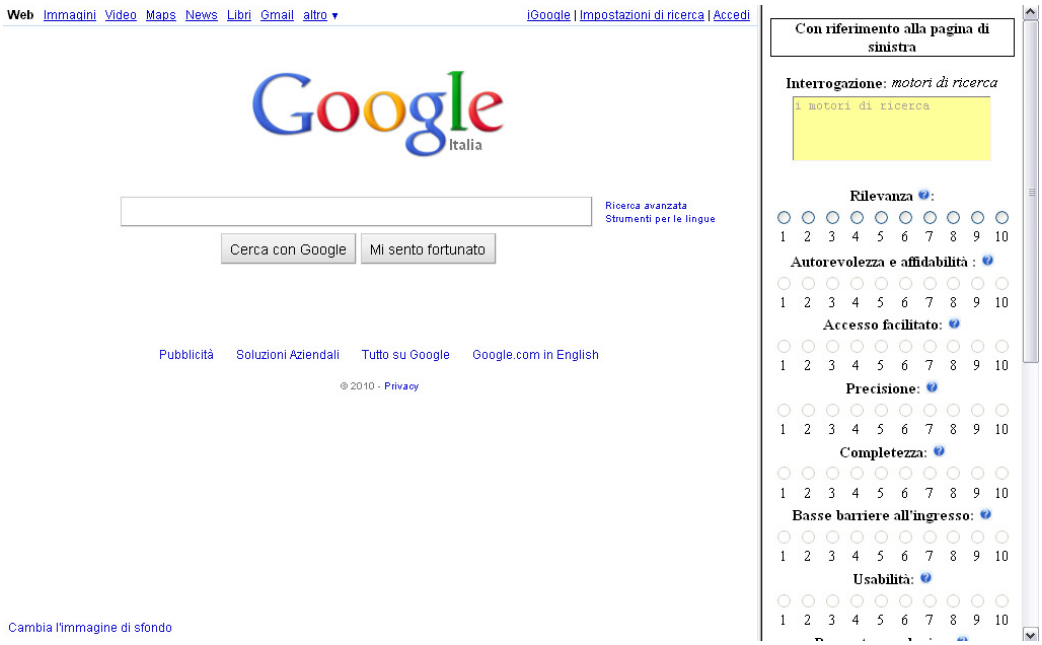

Figura 3.4.: Screenshot della pagina votazioni

4. Campionamento

La base di dati progettata e realizzata come descritto nel capitolo precedente dovrà essere popolata con la pagine web restituite da interrogazioni studiate ad hoc. Dato che non si possono inserire tutte le pagine, dobbiamo procedere con un campionamento e dopo aver effettuato un'analisi statistica per decidere quali pagine inserire e quali no, abbiamo scelto il campionamento a due stadi, dove nel primo stadio si divide la popolazione in grappoli e nel secondo stadio si effettua per ciascuno dei grappoli un campionamento anche diverso tra loro.

4.1. Analisi statistica

Come riportato in [2], per effettuare un'analisi statistica si procede per fasi:

1. Definizione degli obiettivi della ricerca. Gli obiettivi devono individuare le informazioni da ricercare, evitando equivoci, circoscrivendo il territorio e il periodo dell'indagine.
2. Rilevazione dei dati. La rilevazione dei dati può essere completa, quando si esaminano tutti gli elementi oggetto di studio, o parziale, quando ci si limita a studiare un sottoinsieme detto campione. Questa fase è parte del nostro progetto e verrà sviluppata nella sezione 4.4.
3. Elaborazione metodologica. In questa fase si elaborano e si studiano gli elementi restituiti dalla rilevazione dei dati. Nel nostro caso questa fase viene effettuata tramite l'applicazione web descritta nel capitolo precedente.
4. Presentazione ed interpretazione dei dati. Un'accurata illustrazione dei risultati e una disamina particolareggiata delle implicazione operative dei medesimi sono elementi decisivi per il buon esito di un'indagine statistica.
5. Utilizzazione dei risultati della ricerca.

4.2. Elementi di un piano campionario

Definiamo ora gli elementi e la simbologia che nella sezione successiva consentiranno di definire lo schema campionario scelto e di svolgere un'adeguata inferenza statistica.

Indichiamo con $\mathcal{P} = \{U_1, U_2, \ldots, U_N\}$ una popolazione finita di unità statistiche U_i.

L'obiettivo del campionamento da una popolazione finita $\mathcal{P}$ è la selezione di un sottoinsieme $\mathcal{C} \subset \mathcal{P}$, detto campione, la cui dimensione n è inferiore a N .

Il rapporto tra la numerosità n di $\mathcal{C}$ e la numerosità N di $\mathcal{P}$, cioè il numero $\frac{n}{N}$, si chiama frazione di campionamento.

La finalità del campionamento è quella di esaminare le unità statistiche di $\mathcal{C}$ per studiare una variabile X la quale, nella popolazione $\mathcal{P}$, assume i valori $\{X_1, X_2, \ldots, X_N\}$ in corrispondenza di ciascuna unità statistica $\{U_1, U_2, \ldots, U_N\}$.

Ciascuna unità statistica è individuata da un'etichetta che la contraddistingue da tutte le altre, e che assume necessariamente uno dei valori contenuti in $\{1, 2, \ldots, N\}$.

Allora, un campione $\mathcal{C}$ è un sottoinsieme di unità statistiche prescelto da $\mathcal{P}$ per il quale sono note le etichette $\{i_1, i_2, \ldots, i_n\}$ che consentono di identificare gli elementi prescelti. Pertanto, $\mathcal{C} = \{U_{i_1}, U_{i_2}, \ldots, U_{i_n}\}$.

Definiamo spazio campionario $\Omega_n(\mathcal{C})$ l'insieme di tutti i possibili campioni di numerosità n derivabili mediante un prefissato piano di campionamento definito su una popolazione $\mathcal{P}$ di dimensione N.

Un piano di campionamento Π è la definizione di una procedura di selezione di n unità statistiche, per formare il campione $\mathcal{C}$, mediante estrazione da una popolazione finita $\mathcal{P}$ costituita da N unità sulle quali la variabile X è l'oggetto dell'indagine.

Una strategia campionaria consiste nel definire un piano di campionamento.

Lo schema di campionamento, invece, è la procedura operativa mediante la quale si perviene alla effettiva determinazione del campione $\mathcal{C}$ mediante selezione delle unità statistiche dalla popolazione $\mathcal{P}$.

4.3. Indagine campionaria

Nella rilevazione dei dati parziale (o indagine campionaria) la possibilità di limitare la rilevazione ad un insieme di unità di dimensione ben inferiore a quella della popolazione consente di:

- contenere i costi dell'indagine entro limiti accettabili;
- svolgere l'indagine in tempi relativamente brevi;
- raccogliere per ogni unità inclusa nell'indagine un maggior numero di informazioni;
- raccogliere le informazioni con maggior accuratezza.

Sul piano teorico tuttavia l'indagine campionaria presenta due notevoli problemi: il primo, legato al modo in cui deve essere scelto il campione; il secondo relativo ai procedimenti da adottare per estendere l'evidenza campionaria alla popolazione.

L'obiettivo dell'indagine campionaria è quello di descrivere la "realtà" della popolazione alla luce delle osservazioni condotte su un insieme limitato di unità estratte dalla popolazione stessa.

In altri termini il principale obiettivo di un'indagine campionaria è quello di raccogliere dati che consentiranno di generalizzare all'intera popolazione i risultati ottenuti dal campione. Questo processo di generalizzazione è detto inferenza.

4.4. Campionamento a due stadi

Quando la procedura di estrazione di scelta degli elementi di $\mathcal{C}$ da $\mathcal{P}$ secondo il piano Π avviene mediante meccanismi di natura aleatoria si parla di campioni probabilistici, perché di essi si può determinare la probabilità di estrarre una determinata unità statistica. Per questi campioni la variabilità campionaria si può derivare con metodi statistici.

Il piano campionario Π che si adatta meglio al nostro caso è il campionamento a due stadi. Questo piano di campionamento si realizza in due fasi successive di scelta delle unità statistiche nel modo seguente:

- 1° stadio: si divide la popolazione in G grappoli. Un grappolo è un insieme di unità statistiche che sono contigue rispetto ad un criterio logico o naturale. Nel nostro caso quindi le unità nei vari grappoli saranno divise in base alla rilevanza rispetto all'interrogazione.
- 2° stadio: da ciascun grappolo si scelgono le singole unità statistiche secondo un piano di campionamento. Per i vari grappoli è stato adottato il campionamento sistematico in quanto le unità statistiche sono ordinate in base alla loro rilevanza.

È stato scelto questo piano in quanto vogliamo soprattutto i documenti più rilevanti. Si riesce a fare ciò dividendo la popolazione in grappoli secondo la rilevanza e aumentando il passo di campionamento man mano che diminuisce la rilevanza dei documenti contenuti nei grappoli.

4.5. Campionamento sistematico

Sul piano operativo la procedura del campionamento sistematico è molto semplice:

1. le unità della popolazione sono messe in sequenza (nel nostro caso secondo la rilevanza);
2. si associa ad esse un numero da 1 a N;
3. si estrae un numero r casualmente;
4. si seleziona la prima unità considerando l'unità di campionamento associata al numero r;
5. si selezionano le unità successive nel seguente modo:

- partendo da $r+1$ e contando k posizioni si prende l'unità che occupa il posto $r+k$;
- partendo da $r+k+1$ e contando k posizioni si prende l'unità che occupa il posto $r+2k$;
- etc.

Il numero k è detto passo di campionamento ed è uguale all'inverso della frazione di campionamento.

A seconda che la numerosità N della popolazione sia o non sia multipla della numerosità n del campione si ha, rispettivamente, il campionamento sistematico lineare o circolare

Campionamento sistematico lineare Assumiamo che la numerosità della popolazione N sia multipla della numerosità del campione n. Il passo di campionamento è dato da $k = \frac{N}{n}$. Al fine di selezionare un campione di n unità si procede in questo modo:

1. si seleziona casualmente la prima unità estraendo un numero r compreso tra 1 e n;
2. a partire dal numero r si seleziona un'unità ogni k .

In simboli le unità selezionate saranno quelle corrispondenti ai numeri

$$r, r+k, r+k*2, r+k*3, \ldots, r+k*(n-1).$$

Campionamento sistematico circolare Assumiamo che la numerosità della popolazione N non sia multipla della numerosità del campione n. In questo caso il passo di campionamento è dato da $k = \lfloor \frac{N}{n} \rfloor$, dove con $\lfloor \cdot \rfloor$ si intende l'arrotondamento del rapporto $\frac{N}{n}$ al numero intero inferiore.

Si considerano le unità come se fossero in una lista circolare e, al fine di selezionare un campione di n unità si procede in questo modo:

1. si seleziona casualmente la prima unità estraendo un numero r compreso tra 1 e N;
2. a partire dal numero r si seleziona una unità ogni k;
3. se si arriva alla fine della lista si riparte dall'inizio.

Conclusioni

In questa relazione è stata descritta un'applicazione per raccogliere i giudizi di qualità di pagine web e una tecnica per campionare tali pagine restituite da dieci interrogazioni scelte ad hoc.

Nell'applicativo web realizzato le funzioni principali sono la registrazione ed il login nell'home page, la possibilità di scegliere e successivamente di cambiare un'interrogazione e l'inserimento di giudizi di qualità di pagine web da parte degli utenti secondo i seguenti parametri: rilevanza, autorevolezza e affidabilità, accesso facilitato, precisione, completezza, basse barriere all'ingresso, usabilità, presentazione logica, presentazione visiva, lingua, qualità complessiva.

Le pagine web sono state campionate fra quelle restituite dai motori di ricerca sottoponendoli a dieci interrogazioni secondo il campionamento a due stadi. In questo campionamento nel primo stadio è stata divisa la popolazione in G grappoli secondo la rilevanza e nel secondo sono state selezionate le singole unità statistiche tramite il campionamento sistematico. Per campionare soprattutto le pagine più rilevanti è stato aumentato il passo di campionamento man mano che diminuiva la rilevanza dei documenti contenuti nei grappoli.

La fase successiva di questo progetto sarà di applicare gli algoritmi di link analysis partendo dalla struttura topologica del grafo prodotto dopo aver effettuato il crawling del web italiano per determinare i punteggi di qualità. Infine bisognerà confrontare questi punteggi con i giudizi espressi dagli utenti tramite l'applicazione web per desumere se gli algoritmi di link analysis sono efficaci o meno nella selezione delle pagine web di maggiore qualità.

Bibliografia

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2009.

[2] Domenico Piccolo. *Statistica per le decisioni*, il Mulino, Bologna, 2006.

[3] S. Brin and L. Page. *The anatomy of a large scale hypertextual Web search engine.* In Proceedings of the World Wide Web Conference, 1998.

[4] J.M. Kleinberg. *Authoritative sources in a hyperlinked environment.* Journal of the ACM, 46(5):604-632, Sept. 1999.

[5] R. Lempel and S. Moran. *SALSA: The stochastic approach for link-structure analysis.* ACM Transactions on Information Systems, 19(2): 131-160, Apr. 2001.

A. Struttura dell'applicazione web

A.1. Descrizione dei file

index.php Home page del sito che permette di effettuare la registrazione, il login e l'eventuale recupero della password.

iscrizione.php Permette al nuovo utente di iscriversi e controlla la validità dei campi. Una volta terminata l'iscrizione lancia il codice php contenuto in newuser.php.

newuser.php È costituito di solo codice php (la pagina risulta quindi invisibile all'utente) che aggiorna la base di dati con i dati relativi al nuovo utente. Una volta terminata l'operazione porta alla pagina selezioneinterr.php.

selezioneinterr.php Permette all'utente di scegliere e modificare l'interrogazione su cui valutare le pagine. Una volta effettuata la selezione, viene lanciato il codice php contenuto in selezionecheck.php.

selezionecheck.php Come per newuser.php, inserisce o aggiorna la selezione dell'interrogazione da parte dell'utente nella base di dati e lancia direttamente la pagina votazioni.php.

votazioni.php Pagina divisa in due frame. Il primo carica la pagina da valutare (la assegna automaticamente se viene richiesta una nuova pagina, in caso contrario la recupera tra quelle già visualizzate), il secondo carica la pagina menuvotazioni.php.

menuvotazioni.php È inserita nella pagina votazioni e permette di: votare le pagine, scorrere le pagine già valutate ed accedere a nuove pagine da valutare. Inoltre contiene una descrizione dell'interrogazione di cui si stanno valutando i risultati e i link per cambiare interrogazione, modificare il profilo e uscire.

logincheck.php È una pagina di solo codice php che viene richiamata quando, dalla home page, viene effettuato il login. Se il login viene effettuato correttamente porta alla pagina votazioni.php, altrimenti ritorna alla home page.

updateiscrizione.php Permette di modificare i dati dell'account utente. Una volta convalidati i nuovi dati, lancia il codice php contenuto in updateuser.php.

updateuser.php Svolge le stesse funzioni di newuser.php andando però a modificare i dati già esistenti.

db.php Contiene il codice per connettersi alla base di dati.

stile.css Foglio di stile per la formattazione.

A.2. Strumenti utilizzati

DBMS E SERVER: Per la realizzazione della base di dati e delle pagine di applicazione sono stati utilizzati i seguenti strumenti:

- PostgreSQL (vers. 8.4)
- PHP 5
- Apache HTTP Server (vers. 2.2)

In particolare, per poter inviare mail, è necessario aggiornare il file di configurazione di PHP (php.ini) per i seguenti campi:

```
[mail function]
; For Win32 only.
SMTP = out.alice.it (ad esempio)
smtp_port = 25

; For Win32 only.
sendmail_from = xxxyyy@zzz.it
```

B. Documentazione della base di dati

Si vuole realizzare una base di dati che permetta di gestire i giudizi di qualità di pagine web per la quale si vogliono rappresentare gli utenti dell'applicazione, le interrogazioni sottoposte, gli URL da votare, i giudizi espressi, gli algoritmi di link analysis e i punteggi dati da tali algoritmi.

B.1. Requisiti strutturati

Frasi per Utente

Un primo modulo della base di dati consiste nel raccogliere le valutazioni degli utenti su alcune pagine web. Si predisporranno strumenti per la registrazione e la memorizzazione dei vari utenti che parteciperanno alla raccolta, quindi si dovrà implementare un pagina per l'iscrizione di un nuovo utente e una per suoi successivi accessi. Per quanto riguarda la registrazione si dovranno richiedere all' utente i seguenti dati:

Nickname univoco per ogni utente (sarà l'identificatore dell'utente al momento dell'accesso)

Password

Sesso

Nazionalità

Data di nascita

Professione

E-Mail essenzialmente per rimandare la password in caso di perdita

Matricola

Le uniche informazioni obbligatorie sono nickname e password; gli altri campi sono considerati non obbligatori, in particolare la matricola è riservata a quei casi in cui l'utente deve essere identificato come studente. L'utente potrà, in un qualsiasi momento successivo alla registrazione, modificare tali dati.

Frasi per Interrogazione

Si voglia dare l'opportunità di scegliere su quale interrogazione effettuare la procedura di raccolta. Per ogni interrogazione viene messa a disposizione dell'utente l'interrogazione stessa, una breve descrizione sull'argomento trattato e un contatore che misura il numero delle pagine assegnate o già valutate dagli utenti per la specifica interrogazione. Inoltre per ottimizzare le operazioni previste, rappresentiamo anche l'interrogazione corrente per ogni utente, un numero massimo di interrogazioni assegnabili e il numero di interrogazioni effettivamente assegnate per ogni URL agli utenti. È predisposto infine, un ID che identifica univocamente l'interrogazione stessa.

Frasi per Descrizione dei Compiti

Bisogna informare l'utente su quali siano i suoi compiti e come si svolgerà la valutazione delle pagine descrivendo le varie voci che compongono una valutazione, dicendo che è richiesta una conoscenza degli argomenti, etc...

Frasi per Soglia

Una serie di parametri utili per la gestione web della base di dati.

Frasi per Restituzione

Per il voto dell'utente si voglia memorizzare un primo indice chiamato rilevanza

Frasi per Parametro di Qualità

Per ogni Parametro di qualità è di interesse memorizzare il nome, una descrizione e un ID che lo identifica univocamente. Ha senso esprimere un voto per il parametro di qualità se la rilevanza relativa è maggiore o uguale a 6.

Frasi per Indirizzo Web (URL)

Per quanto riguarda l'indirizzo della pagina in questione è di interesse mantenere in memoria l'indirizzo stesso e un ID più agevole che lo identifica univocamente.

Frasi per Algoritmo di Ricerca

Per ogni Algoritmo di Ricerca preso in considerazione rappresentiamo un nome, una descrizione testuale e un ID alfa-numerico che lo identifica univocamente.

Frasi per QualityScore

Ogni algoritmo di Ricerca emette un punteggio per ogni URL o interrogazione e si assume che possa valutare più URL e interrogazioni. Perciò per ogni QualityScore rappresentiamo un valore numerico e un ID alfa-numerico che lo identifica univocamente.

Una terza utilità da implementare è quello che dà un resoconto sommario e veloce della quantità di dati che sono stati finora raccolti per avere un'idea di quanto materiale si dispone nella base di dati.

Operazioni supportate

1. Visualizzazione della descrizione del progetto
2. Iscrizione utente e aggiornamento estremi
3. Login ed eventuale recupero password (invio per email)
4. Scelta dell'interrogazione
5. Visualizzazione e aggiornamento voti delle pagine
6. Visualizzazione delle interrogazioni invocate fino ad oggi

B.2. Schema ER

Viene proposto uno schema ER per il progetto:

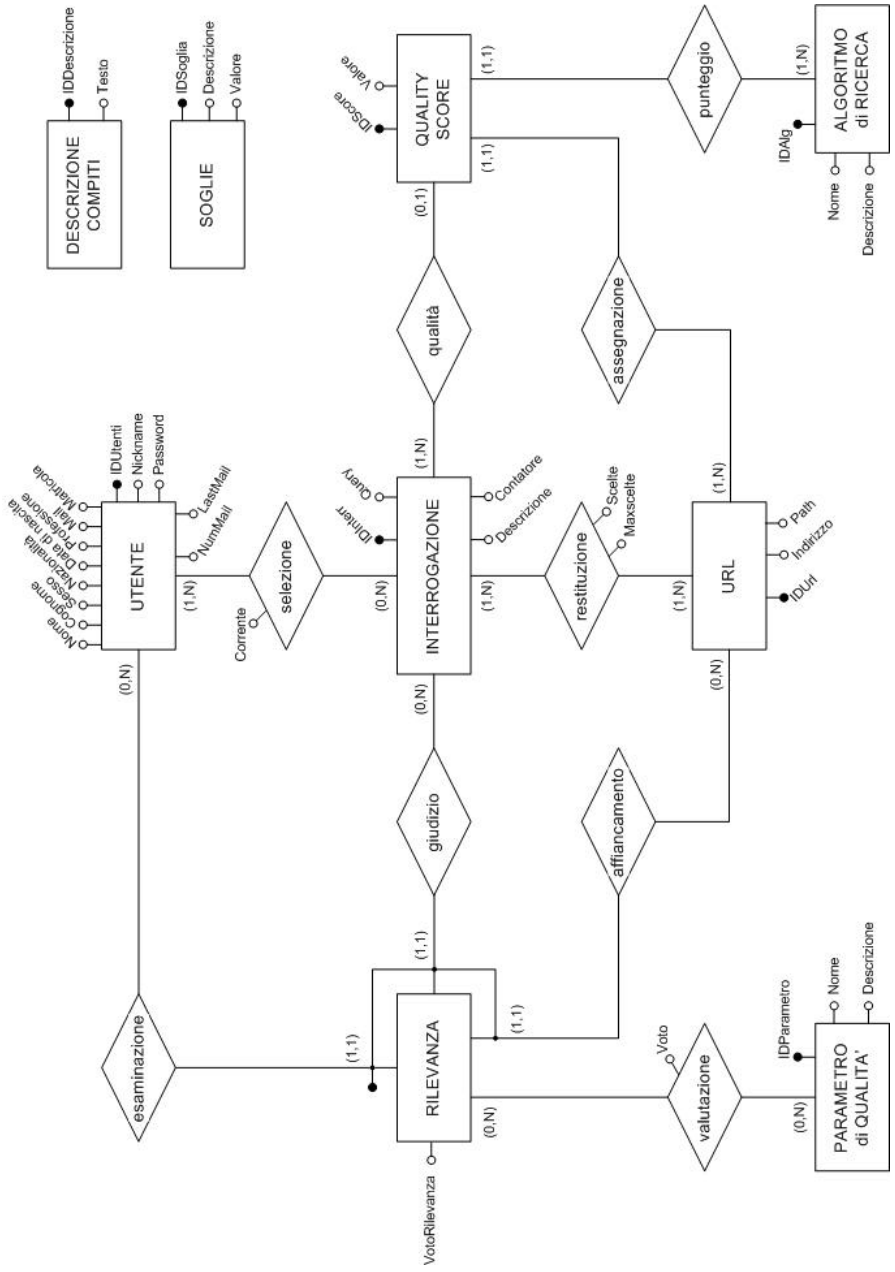

Figura B.1.: Schema ER della base di dati

Descrizione delle entità

UTENTE modella gli utenti mediante i seguenti attributi

- IDUtenti
- Nickname
- Password
- Nome
- Cognome
- Sesso
- Nazionalità
- Data di nascita
- Professione
- E-Mail
- Matricola
- NumMail
- LastMail

INTERROGAZIONE modella le interrogazioni che vengono considerate nel progetto

- IDInterr
- Query
- Descrizione
- Contatore (numero di pagine assegnate o già valutate dagli utenti per l'interrogazione in questione)

URL modella le pagine che verranno proposte agli utenti

- IDUrl
- Indirizzo

RILEVANZA modella le pagine analizzate dagli utenti e associa un voto di rilevanza

- VotoRilevanza

PARAMETRO QUALITA modella i singoli parametri sui quali viene sviluppata la votazione delle pagine

- IDParametro
- Nome
- Descrizione

QUALITY SCORE modella i valori generati dagli algoritmi di qualità

- IDScore
- Valore

ALGORITMO DI RICERCA modella gli algoritmi di ricerca che vengono studiati nel progetto

- IDAlg
- Nome
- Descrizione

DESCRIZIONE COMPITI memorizza la descrizione dei compiti che un utente deve effettuare (inserita nella base di dati solo per ragioni di comodità)

- IDDescrizione
- Testo

SOGLIE memorizza dei valori numerici utili per la gestione delle pagine da analizzare

- IDSoglia
- Descrizione
- Valore

Descrizione di Associazioni particolari

Nello schema presentato sono presenti tre associazioni che meritano un'attenzione particolare:

- La prima è quella che lega l'utente alle interrogazioni scelte tra le quali viene segnata l'ultima selezione (cioè l'interrogazione corrente per ciascun utente).
- La seconda è data dalle associazioni tra interrogazioni e URL corrispondenti, per ciascuna delle quali viene definito un numero massimo di assegnazioni agli utenti (MaxScelte) e viene aggiornato il numero di assegnazioni effettivo (Scelte).
- Infine le restituzioni vengono associate ai diversi parametri di qualità stabiliti e, per ciascun accoppiamento che si crea al momento della votazione, viene memorizzato il voto di tale restituzione per tale parametro.

B.3. Schema logico

Lo schema logico-relazionale della base di dati, porta ad una struttura come quella illustrata in figura B.2, in cui vengono rappresentate le entità descritte in precedenza sotto forma di tabelle di una base di dati. Inoltre vengono evidenziate le relazioni che intercorrono tra le diverse entità.

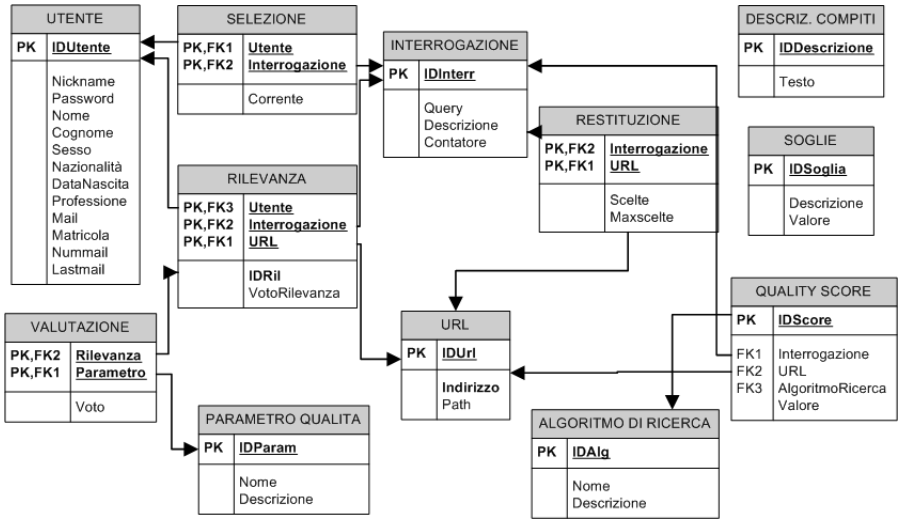

Figura B.2.: Schema logico della base di dati

Tra UTENTE e INTERROGAZIONE è presente un'associazione molti-a-molti, infatti si vuole dare l'opportunità agli utenti di scegliere più di una interrogazione da analizzare, eventualmente controllando che prima di fornire tale opportunità, sia terminata l'analisi dell'interrogazione precedente. Questa relazione è rappresentata dalla tabella SELEZIONE che unisce, appunto, INTERROGAZIONE e UTENTE.

Le interrogazioni al motore di ricerca restituiscono una serie di risultati che vengono inseriti nella relazione URL senza duplicati nel caso in cui lo stesso indirizzo si trovi tra i risultati di più interrogazioni. Questa possibilità rimane comunque da non escludere, perciò, come indicato dallo schema ER, l'associazione tra INTERROGAZIONE e URL è ancora una volta di tipo molti-a-molti e viene realizzata attraverso la tabella RESTITUZIONE la quale rappresenta realmente ciascuna coppia interrogazione-risultato ottenuta attraverso le interrogazioni tramite motore di ricerca.

Gli URL ottenuti dalle ricerche, vengono poi assegnate agli utenti per l'analisi. Questa assegnazione è rappresentata dall'inserimento di un dato all'interno della tabella RILEVANZA che per tale pagina, rispetto a una data interrogazione, memorizza il voto di rilevanza dato dall'utente. Per comodità, la relazione Rilevanza è indicizzata dal contatore IDRil. È per tan-

to opportuno includere un vincolo di non nullità e univocità per quanto riguarda l'attributo IDRil.

Operazioni sulla base di dati

Nell'inserimento e nell'aggiornamento dei dati della base di dati è necessario tenere in considerazione i vincoli che legano le diverse entità presenti. A livello di programmazione, questi vincoli si tradurranno in controlli sui dati già presenti nella base di dati che rispettino le specifiche di progetto.

Raggruppiamo tali operazioni secondo l'ordine temporale stimato di esecuzione:

Operazioni_preliminari

- Inserimento delle interrogazioni, indicizzate dall'attributo chiave IDInterr, con Contatore inizializzato a 0.
- Inserimento dei parametri di qualità con cui si valuteranno le pagine.
- Inserimento degli algoritmi di ricerca da testare sulle pagine.
- Inserimento delle descrizioni dei compiti: spiegazioni sul progetto e sulle operazioni che l'utente deve eseguire che verranno richiamate nelle opportune pagine web del sito.
- Definizione delle soglie:
  1. soglia per il contatore delle interrogazioni cioè il numero massimo delle restituzioni per ciascuna interrogazione;
  2. soglia per il numero di risultati da rilevare tramite motore di ricerca per pagina;
  3. soglia per il voto di rilevanza, al di sopra della quale una pagina può essere valutata tramite i parametri di qualità perché ritenuta, appunto, rilevante.

Recupero_delle_pagine_di_una_data_interrogazione

- Recupero degli indirizzi da un motore di ricerca interrogato secondo l'interrogazione data (il numero di risultati da tenere in considerazione sono dati dalla soglia 2).
- Inserimento degli indirizzi ottenuti nella tabella URL e aggiornamento della relazione RESTITUZIONE (inizializzando il contatore Scelte a zero e la soglia MaxScelte al valore desiderato).

Registrazione dell'utente

- Creazione di un nuovo utente secondo le informazioni fornite.
- All'utente vengono mostrate le interrogazioni ancora disponibili, cioè tutte le interrogazioni il cui contatore non supera la soglia 1.

- La scelta dell'interrogazione da parte dell'utente implica l'aggiornamento della relazione UTENTI-INTERROGAZIONI, cioè l'inserimento della coppia IDUtenti-IDInterr nella tabella SELEZIONI.

Valutazione delle pagine da parte dell'utente

- All'utente che ha scelto una data interrogazione, viene presentata una prima pagina tra quelle disponibili tenendo conto del controllo sulla relazione RESTITUZIONE: Scelte minore di MaxScelte.
- Si inserisce una rilevanza, nell'omonima tabella.
- Quando l'utente esprime la sua valutazione sulla rilevanza di una pagina proposta, si aggiorna il rispettivo attributo di RILEVANZA e si incrementa il contatore relativo all'interrogazione in questione.
- Se la pagina è rilevante, cioè il voto dato è superiore al voto di soglia 4, l'utente può passare alla votazione di qualità inserendo per ogni voto dato dall'utente, una coppia Rilevanza-Parametro nella tabella VALUTAZIONE con il voto specifico.
- Si vuole lasciare all'utente la possibilità di modificare in ogni momento i voti di rilevanza e sui parametri di qualità. Sarà necessario eliminare dalla relazione VALUTAZIONE i dati relativi.
- L'utente può in ogni momento scegliere di passare a una pagina successiva oppure può decidere di cambiare interrogazione. Si dovrà pertanto aggiungere nella tabella SELEZIONE la nuova scelta, aggiornando opportunamente l'attributo Corrente.

B.4. Codice SQL delle tabelle

```
CREATE TYPE genere AS ENUM ('M', 'F');

CREATE TABLE UTENTE (IDUtente integer PRIMARY KEY, Nickname varchar
    (50), Password varchar(20) NOT NULL, Nome varchar(50), Cognome
    varchar(50), Sesso genere, Nazionalita varchar(20), DataNascita
    date, Professione varchar(50), Mail varchar(50), Matricola
    integer, NumMail smallint DEFAULT 0, LastMail date);

CREATE TABLE INTERROGAZIONE (idinterr integer PRIMARY KEY, Query
    varchar(50) NOT NULL, Descrizione text, Contatore integer
    DEFAULT 0);

CREATE TABLE SELEZIONE (utente integer REFERENCES UTENTE (IDutente)
    , interrogazione integer REFERENCES INTERROGAZIONE (idinterr),
```

```
    Corrente BOOLEAN DEFAULT FALSE, PRIMARY KEY (utente,
    interrogazione));

CREATE TABLE URL (IDUrl integer PRIMARY KEY, Indirizzo character
    varying NOT NULL, path character varying);

CREATE TABLE RESTITUZIONE (interrogazione integer REFERENCES
    INTERROGAZIONE (idinterr), Url integer REFERENCES URL (IDUrl),
    scelte smallint DEFAULT 0, maxscelte smallint DEFAULT 1, PRIMARY
     KEY (interrogazione,url));

CREATE TABLE RILEVANZA (utente integer REFERENCES UTENTE (IDutente)
    , interrogazione integer REFERENCES INTERROGAZIONE (idinterr),
    url integer REFERENCES URL (IDUrl), idril integer UNIQUE,
    votorilevanza smallint DEFAULT -1, PRIMARY KEY (utente,
    interrogazione, url));

CREATE TABLE PARAMETROQUALITA (idparam integer PRIMARY KEY, Nome
    varchar(20), Descrizione text);

CREATE TABLE VALUTAZIONE (rilevanza integer REFERENCES RILEVANZA (
    idril), parametro integer REFERENCES PARAMETROQUALITA (idparam),
     Voto smallint, PRIMARY KEY (rilevanza, parametro));

CREATE TABLE ALGORITMODIRICERCA (IDAlg integer PRIMARY KEY, Nome
    varchar(20), Descrizione text);

CREATE TABLE QUALITYSCORE (IDScore integer PRIMARY KEY,
    interrogazione integer REFERENCES INTERROGAZIONE (idinterr), url
     integer REFERENCES URL (IDUrl), AlgoritmoRicerca integer
    REFERENCES algoritmodiricerca(IDAlg), Valore smallint);

CREATE TABLE DESCRIZCOMPITI (IDDescrizione integer PRIMARY KEY,
    Testo text);

CREATE TABLE SOGLIE (IDSoglia integer PRIMARY KEY, Descrizione text
    , Valore integer);

GRANT select, update, delete, insert on UTENTE, INTERROGAZIONE,
    SELEZIONE, URL, RESTITUZIONE, RILEVANZA, VALUTAZIONE,
    QUALITYSCORE, SOGLIE to webdb;
```